"HER2 immunohistochemistry inter-observer reproducibility in 205 cases of invasive breast carcinoma additionally tested by ISH" Answer to the statistical issue to avoid misinterpretation☆

Carlos Casterá Redal*, Laia Bernet Vegué

Hospital Universitario de la Ribera, Crta Corbera km 1, 46600 Alzira, Valencia, Spain

ARTICLE INFO

Keywords:
HER2 assessment
Inter-observer reproducibility
Major consensus hypothesis
Equivocal cases
Invasive breast carcinoma

ABSTRACT

Assessment of HER2 biomarker in invasive breast carcinoma patients allows a specific therapeutic approach. Clinical guidelines indicate immunohistochemistry (IHC) and in situ hybridization (ISH) to test HER2, however both have drawbacks which results in low reproducibility of results especially in equivocal cases. Our main objective is to quantify inter-observer IHC reproducibility and cross it with the ISH result. Our series includes 205 invasive breast carcinoma cases sent for ISH retest from 14 hospitals, 5 observers to assess the IHC and 1 observer for ISH of each case. We found that the observers only achieve an absolute agreement for IHC in 1 out of 3 cases. The inter-observer concordance for IHC is low ($0.2 \leq k \leq 0.4$) or moderate ($0.41 \leq k \leq 0.6$). In ISH positive cases the concordance for IHC is higher than in the ISH negative cases. In conclusion, the study shows low and moderate IHC inter-observer concordance, finding the more worrying values among the ISH negative cases which are the most part of this particular sample. Subjective interpretation of the techniques, among other factors, has negative impact in HER2 evaluation. To offset this limitation we have checked that reaching a consensus from different observers for HER2 IHC assessment improves the results.

We'd like to express our appreciation for receiving this letter, which faces some methodological observations from our article "HER2 immunohistochemistry inter-observer reproducibility in 205 cases of invasive breast carcinoma additionally tested by ISH" recently published in this journal.

We appreciate very much the comments on our work, because this is a learning opportunity for us and all the readers of the journal.

First of all, regarding the kappa limitations, we were aware of them and agree that perhaps the application of the weighted kappa would have been a better option to evaluate the inter-observer agreement in our study. Furthermore, we calculated it and obtained slightly higher values of agreement, but the fact of being more familiar with kappa, and its more frequent use favored our decision of using it. However, all these methodological issues, although a key nuance for the statistical interpretation of the results, do not modify the conclusions of our work and some relevant statements.

For example (paper quotes) [1]:

- In conclusion, the most surprising fact regarding the reproducibility among HER2 IHC observers in this particular sampling (Fig. 1)

indicates that in only about 1 out of 3 cases (31.22%) is there a total agreement among the 5 observers.

- We know that practice and experience in HER2 evaluation leads to more robust results, so an optimal solution - especially in the case of biomarkers/therapeutics - would be to have the diagnosis performed in large, accredited, EQA laboratories by well-trained specialists with state-of-the-art equipment and software.
- Finally, in this study we found that many factors (subjectivity, observers, sampling, diagnostic algorithm and others) influence the IHC concordance values among observers.

Secondly, one of our paper's conclusions indeed is that subjective interpretation of test results, among other factors, has a negative impact on HER2 evaluation. Although kappa is not the most adequate statistic to support this idea, in our daily practice we have found that, unlike mathematics, diagnosis is not an exact science. Diagnostics tries to identify the nature or biology of diseases, and when a biomarker is evaluated by different techniques we find a plethora of possible outcomes depending on many variables (pre-analytical, analytical and post-analytical), so when evaluating HER2 we find a wide range of
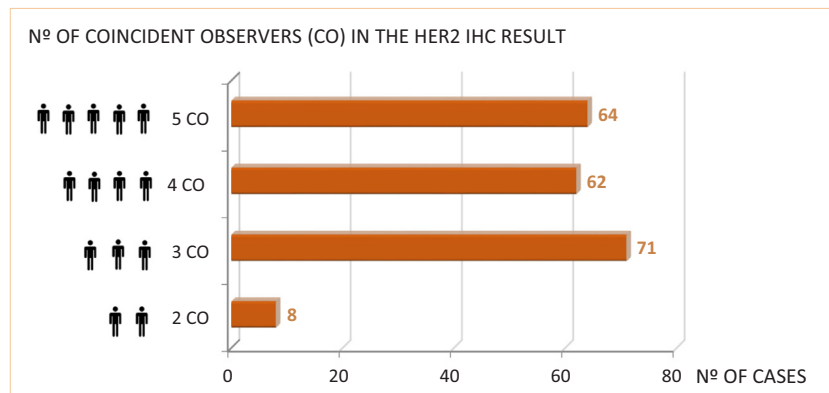

Fig. 1. Number of cases distributed according to the number of observers that agree (Coincident Observers; CO) in the HER2 IHC result.

grays and shades of color in the process of determining the final result. At this point, we consider that our goal should be to make the diagnoses as accurate, precise and reproducible as possible, so we must try to reduce the impact of all factors that prevent it - such as human subjectivity in the interpretation of the results of Immunohistochemistry (IHC) - and enhance all other factors that may allow it.

Finally, I would like to mention my colleagues, who have made a great contribution with their comments on the limitations of the kappa statistic to assess agreement.

References

[1] Casterá C, Bernet L. HER2 immunohistochemistry inter-observer reproducibility in 205 cases of invasive breast carcinoma additionally tested by ISH. Ann Diagn Pathol 2020 Apr;45:151451https://doi.org/10.1016/j.anndiagpath.2019.151451.